Video Augmentation for Improving Audio Speech Recognition under Noise

Samuel Pachoud Shaogang Gong Andrea Cavallaro
Queen Mary, University of London
London, United Kingdom
{spachoud,sgg}@dcs.qmul.ac.uk
andrea.cavallaro@elec.qmul.ac.uk

Abstract

For the recognition of speech, in particular spoken digits, captured in video with poor sound due to noise, we develop a novel audio-visual fusion technique that performs significantly better than utilising either audio or video signal alone. Specifically, we present an audio-visual intermediate fusion strategy to locate speaker dependant pronounced digits in continuous video recorded with sound. A model template for each digit is represented in a single audio-visual feature space using a set of spatio-temporal visual features at multiple scales together with a set of thirteen Mel Frequency Cepstral Coefficients as audio features. Using a unified structure for both visual and audio feature selection and extraction, we solve the problem of one-to-one correspondence between the audio and visual spaces caused by differences in data sampling rates. To combine the two modalities, we adopt an intermediate fusion strategy by combining the two modalities in a probabilistic sequence matching function, permitting automatic segmentation of a continuous probe video sequence and matching with available model templates. For experiments, the CUAVE [17] database was used to compare our scheme with two alternative methods. The evaluation shows that the proposed approach outperforms the others both in recognition accuracy and robustness in coping with variations in probe sequences.

1 Introduction

For perceiving facial emotion and behaviour, humans combine the acoustic waveform (audio information) and the movements of the lips, tongue and other facial muscles (visual information) generated by a speaker. The McGurk effect [12] establishes this bimodal speech perception by showing that, when conflicting audio and visual stimuli are presented to an individual, the latter may assimilate a new stimulus, different from the other two. This implies an increase of the importance of the visual information, especially in noisy environments, quantified by Sumby and Pollack [21]. Such observations have motivated interest in developing systems for automatic recognition of audio-visual speech. To that end, a fundamental challenge is how to combine effectively audio and visual signals

in such a way that visual information can assist audio especially when its signal-noise ratio is low.

In this work we aim to address the issue of recognising spoken digits in a noisy environment with the aid of visual input. We consider a intermediate fusion strategy and we address the problem of synchronisation between audio and visual signals. A linear interpolation based approach adopted by existing techniques [9, 23] does not secure sufficient one-to-one correspondences and risks in inserting errors. Instead, we consider constructing a similar structure for both audio and visual feature selection and extraction. As speech perception and audio features are time related, we adopt a space-time volume based feature representation based on our previous work [16]. Then instead of using a single joint audio-visual feature, we formulate an intermediate fusion strategy, which uses models that infer the synchrony at the phoneme or word boundaries, to perform automatic audio-visual speech recognition under significant audio noise. Our approach aims to improve recognition rate compared to that of using either visual information alone or audio alone, especially with poor sound.

Our aim is to build a set of model templates consisting of a database of all the visemes[1] of a studied language. These templates can then be deployed to provide a concise and generative representation at an atomic level, e.g. the English language is composed of only about fifteen visemes. To begin, we focus in this paper on building model templates for automatically segmenting and recognising 10 digits appearing randomly in continuous probe video sequences with spoken sound.

2 Related Work

There are two fusion strategies that have been adopted by existing techniques aiming to combine audio and visual modalities: feature space fusion where a single feature space is constructed by concatenating audio and visual features; decision level fusion where separate recognisers are trained for the two modalities before a joint likelihood function is designed for a final decision making. In particular, early stage fusion in feature space is used in [4, 8, 9, 14, 23] and is optimal when the modalities are highly correlated. Three major techniques are used to perform feature fusion: Hidden Markov Model (HMM), Gaussian Mixture Model (GMM) and Coinertia Analysis (COIA). One of the problems in feature fusion strategy is to retain the one-to-one correspondences between the audio and the video signal, which have often different sampling rates. Kaynak *et al.* [9] use a four dimensional visual geometric feature vectors like outer-lip or inner-lip parameters and thirteen Mel Frequency Ceptsral Coefficients (MFCC, audio feature coefficients). The synchronization between the two modalities is assured by sampling the visual features with a low-pass interpolation. Then they concatenate the two feature vectors to form a sixteen dimensional joint feature vector, which is used to train six different HMMs with four, five states and eight, sixteen and thirty-two Gaussian mixtures for each word. Chen [4] extracts three visual features (the mouth width, the height of the upper lip, and the height of the lower lip) and converts each time window into sixteenth-order Linear Prediction Coding (LPC) coefficients. The cascaded and weighted audio-visual feature vector fed either a GMM or a HMM. A distribution $p_{va(video,audio)}$ is modelled as a GMM

[1] A viseme is a basic unit of speech in the visual domain that corresponds to phoneme (which is the basic unit of speech in the acoustic domain).

to then estimate the conditional expectation of video given audio, i.e. $E[video|audio]$. In the second case, each word is trained in a five states and three Gaussian mixtures HMM. Active Appearance Model (AAM) [5] is used to provide the visual feature vectors in [23]. Twelve MFCCs are used to form a twenty-two dimensional audio-visual space. A linear interpolation is applied to the visual appearance parameters to reduce the dimension and to guarantee the correspondence. Then the authors apply a HMM or GMM to calculate the joint probability of the audio-visual feature space distribution. Nefian *et al.* [14] create a fifteen coefficients visual feature vector using a LDA and a 2D-DCT and form a joint audio-visual vector with thirteen MFCCs. Then the authors compare two types of dynamic Bayesian networks, the factorial and the coupled HMM. They conclude that the coupled HMM performs better than the factorial one. Finally a Coinertia Analysis (COIA) is applied in [8], in which a joint audio-visual feature vector with thirteen MFCCs is formed. Their delta and delta-delta parameters are concatenated with visual feature vector derived from mouth colour information and geometric features. To ensure the synchronization between the audio and video signals, the latter is resampled to a higher frequency. COIA is then used to model linear combinations between the two modalities. The major issue, with the methods explained above, is the one-to-one correspondence between the audio and visual spaces. Using interpolation of the video signal, some useful and discriminative information are lost. On the other hand, late stage fusion algorithms [1, 10] utilise the two single-modality classifier outputs, often multi-stream HMMs, to recognize audio-visual speech. In [1], the visual feature vector consists of projections weights (snake of the lip contours) and the first and second order derivative, while the audio feature vector contains twelve MFCCs and the first and second order derivative too. To match the sampling rate , the visual features are interpolated. Then using a multi-stream HMMs, audio and visual log-likelihoods are combined using weights that capture their reliability. Decision fusion strategies are, by definition, pretty straightforward however they are highly dependant on the respective weights for the audio and the visual modalities, which can become arduous when the audio signal is very noisy. Bringing those two strategies together, Sargin *et al.* [20] combine decision and feature fusion following canonical correlation analysis. This method allows them to manage the synchronisation and the fusion of the two modalities. A summary of lip-reading approaches is shown in Table 1.

3 Sequence matching for digits recognition

For recognising spoken digits through audio (speech) and video (lip movements) information, we propose here an audio-visual extension of the visual-only system reported in [16]. Their system is based on extracting a set of video model digit templates representing digits 0 to 9 separately, before matching any probe video sequence against these model templates. In a probe sequence, the order and the number of pronounced digits are unknown. In contrast, our model consists of two major parts: (1) audio and visual feature selection and extraction, (2) intermediate fusion strategy by combining audio and video features in a probabilistic sequence matching function. Our first step automatically defines and extracts sets of audio and visual features without any manual labelling of feature points, alignment between frames and samples, or scale normalisation in space or in time.

Ref	Segmentation	Feature extraction		Fusion
		audio	visual	
				Decision
[10]	MESH / DFT	HMM	PCA / LDA	-
[1]	-	MFCC	Snake and parabolas	HMM
				Feature
[9]	-	MFCC	height and width, area and angle	HMM
[14]	colour information	MFCC	DCT / LDA	modified HMM
[4]	GMM	LPC	1 width and 2 heights	GMM/HMM
[23]	-	MFCC and PCA	AAM	GMM/HMM
[8]	-	MFCC	colour and geometric features	COIA
				Decision/Feature
[18, 19]	DCT/DTW	MFCC	LDA/MLLT	HMM
[22]	-	MFCC	DCT	HMM
[20]	-	MFCC	DCT	CCA

Table 1: A summary of audio-visual speech reading approaches divided into two main groups. AAM: Active Appearance Model; CCA: Canonical Correlation Analysis; COIA: Coinertia Analysis; DCT: Discrete Cosine Transform; DFT: Discrete Fourier Transform; GMM: Gaussian Mixture Model; HMM: Hidden Markov Model LDA: Linear Discriminant Analysis; LPC: Linear Predictive Coding; MESH: Collection of vertices and polygons; MFCC: Mel Frequency Cepstral Coefficients; MLLT: Maximum Likelihood Data Rotation; PCA: Principal Component Analysis.

The features, which are referred to as macro-cuboïds [2] (MC), are defined in space and over time. In the visual channel, the macro-cuboïds are then divided into a set of cuboïds, covering at least some parts of the lip movement. These cuboïds are represented at multiple spatial scales. In the audio channel, thirteen Mel Frequency Cepstral Coefficients are computed per time-window[3] (TW). The number and the length of the time-windows correspond to the number and the scale over time of the macro-cuboïds, hence we obtain a similar dimension for the two modalities. This approach allows us to have the same structure for the audio and the visual feature extraction, which solves the synchronisation issue explained in Section 1. Then a kernel-based maximum likelihood matching function is utilised to find the best match of all the macro-cuboïd candidates in a probe sequence for a model template. Digit recognition is determined by a histogram computed with the highest probability of a model macro-cuboïd (i.e. the biggest bin) indicating both the existence of a digit and its exact location in the probe sequence. Figure 1 gives an overview of our approach. We shall describe the details in the following.

3.1 Audio-visual feature selection and extraction

For visual features, instead of extracting the principal components of lip movement in order to establish a one-to-one correspondence between phonemes of speech and visemes

[2]The term macro-cuboïd comes as a spatio-temporal extension of macroblock (16x16 pixels are used for motion estimation and compensation in traditional video encoders liker H.261 or MPEG-1/2), which is a widely used term in video compression.

[3]The term time-window comes from the window function principle which is a function that is zero-valued outside of some chosen interval, which is a widely used term in signal processing.

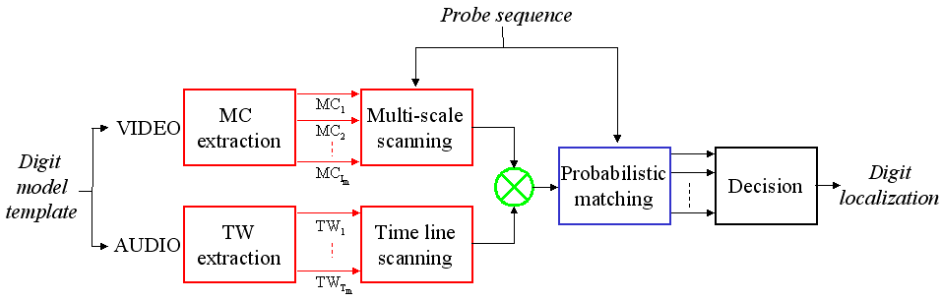


Figure 1: Processing blocks of our digit recognition system. Each digit model template is represented by a set of spatio-temporal visual macro-cuboïds (I_m macro-cuboïds) and of acoustic time-windows (T_m time-windows). Then for each macro-cuboïd and time-windows, we perform a feature difference through the probe sequence. A probabilistic matching function is computed for each scan. Then a histogram-based decision is made to localise and recognise the model template.

of lip shape [19, 8, 14], we consider comparing the movements of lips generated by a speaker (*probe* movements) with the movements of lips of particular words (digits only in this work) in a certain language (*model* movements). This consideration induces space-time features, which embed the lip movements. The idea of working in space and over time is exploited in [2, 3, 6, 15, 16]. These approaches are based either on matching space-time trajectories of moving regions or on detection of interest points (features) within a stack of frames. Such an approach is in strong contrast to the more traditional matching of explicit landmark interest points (e.g. corners, edges), which is the basis of most existing image-to-image matching techniques [9, 11]. More information about the selection and extraction of the visual features is presented in [16].

For audio features, cepstral features are very widely used in audio speech recognition systems. Here we use a variant of the standard cepstrum, the Mel Frequency Cepstral Coefficients (MFCCs). To obtain MFCCs, a windowing function, the popular Hamming window in our case, is applied on the speech signal before the short-term log-power spectrum is computed (using Discrete Fourier Transform). Then the spectrum is wrapped along its frequency axis f into the mel-frequency axis. This is to approximately reflect the human's ear perception. Then the resulting wrapped power spectrum is convolved typically by a blank of triangular filters (between 30 and 40 filters). The latter is approximately linear below 1kHz and logarithmic above 1 kHz; the mel scale effectively reduces the contribution of higher frequencies to the recognition. Finally, the MFCCs are obtained by computing the DCT using

$$MFCC(n) = \sum_{m=1}^{M} X_m cos\left[n(m-0.5)\frac{\pi}{M}\right] \qquad \text{n=1,...N} \qquad (1)$$

where n is the index for cepstral coefficients and m is the index for filters. X_m is the signal after being convolved by the Mel filter bank.

In many automatic speech recognition system, the 0^{th} coefficient of the MFCC is ignored due to its unreliability. In [7], they demonstrate that the 0^{th} coefficient is regarded

as a collection of average energies of each frequency bands in the audio signal. The common number of coefficients N is thirteen.

During the feature selection and extraction process, a set of model digit templates is divided into several macro-cuboïds (MC) and time-windows (TW), which are automatically selected to cover the whole space and time of the model digit templates (exhaustive division). The division of the model templates is based on the video modality (macro-cuboïd). Then the audio speech is divided in time-windows according to the scale over time of the macro-cuboïds, hence the same length in time of the TW than the MC. Following the feature selection and extraction, the matching between a model template and a probe sequence requires the computation of a probability function between the extracted macro-cuboïds and the corresponding time-window from each model within the probe sequence. In the visual channel for each model template, this operation is performed I_m times (see Figure 1), where I_m corresponds to the number of model macro-cuboïds over multiple scales of the m^{th} model template. Each model template, in the audio channel, is divided in T_m time-windows to cover the whole model template. One T_m time-window is always coupled with several macro-cuboïds as there are more than one *MC* per time scale.

3.2 Feature fusion and probabilistic sequence matching

After the selection and the extraction of the audio and visual features, a probabilistic sequence matching is performed. The probability of a model macro-cuboïd and its respective time-windows ($P(\mathrm{AV})$) to be matched with the probe sequence, *PS*, is as follows:

$$P(\mathrm{AV}_i^m, \mathrm{PS}) = \frac{1}{\sigma_d^2 + \sigma_l^2 + \sigma_D} \prod_i^N e^{-\frac{|\Delta d_i|^2}{2\sigma_d^2}} e^{-\frac{|\Delta l_i|^2}{2\sigma_l^2}} e^{-\frac{D(MTW,PS)}{\sigma_{D(MTW,PS)}}} \tag{2}$$

The first two exponential terms represent the visual modality and the last one represents the audio channel. Δd_i and Δl_i are respectively the differences and local displacements between descriptors of the cuboïds C_j^m and their correspondents in the probe sequence. σ_l is equal to the norm of the diagonal of macro-cuboïds MC_i^m. σ_d is determined empirically to give an equivalent weight of $-\frac{|\Delta d_i|^2}{2\sigma_d^2}$ with $-\frac{|\Delta l_i|^2}{2\sigma_d^2}$ in Equation (2). An explanation of the descriptors is given in [16].

The optimal alignment $D(MTW,PS)$ between the two MFCCs vectors in the model time-window (MTW) and the probe one is computed following the Dynamic Time Warping (DTW) principle [13]: at first the sum of local distances between the elements of the two MFCC vectors $d(MTW_i, PS_j)$ is computed. Then the algorithm recursively calculates the optimal alignment $D(MTW,PS)$ for each point element while confirming to local constraints (heuristics) regarding how an optimum alignment warp reaches that point. In our approach, we use the following local constraints of each element (i, j):

$$\tilde{D}(i,j) = d(i,j) + min\left\{\tilde{D}(i-1,j), \tilde{D}(i-1,j-1) + d(i,j), \tilde{D}(i,j-1)\right\} \tag{3}$$

where $i = 1,\ldots,N$ and $j = 1,\ldots,M$. Finally the optimal alignment is equal to:

$$D(MTW,PS) = \tilde{D}(I,J)/N(g) \tag{4}$$

where $N(g)$ is the path normalization factor , which allows comparison between different warps. In Equation (2), $\sigma_{D(MTW,PS)}$ is the variance of the sum of local distances between the elements of the two MFCC vectors $d(MTW_i, PS_j)$.

3.3 Digit recognition by lip-reading

To allow the digit localisation and recognition, an histogram of the i^{th} joint model macro-cuboïds and time-windows with the highest probability to be in the probe sequence is computed. The biggest bin indicates the position of the most likely match between a model digit template and a segment in a probe sequence. This information gives us the recognition and the localisation of digits in the probe sequence. If we assume that a set of model templates fully represents a language, then each part of a probe sequence can be decrypted. The model templates will consist of a database of all the visemes of the language. The main advantage of this is that the database will be concise and generative, because for instance, as mentioned earlier, the English language is composed of fifteen visemes only. For the examples used in this paper, we need to model 10 digits only in order to analyse any arbitrary combination of pronounced digits in a video sequence.

4 Experiments

For our experiments, we use the CUAVE database [17]. The CUAVE corpus is a moving-talker speaker-independent database, designed to support research into audio-visual speech recognition.

The database is converted into grey-level images and each frame is cropped around the mouth (ROI). We divided the dataset into two parts: one part is used to generate the model templates and the other is used for the probe sequences. Each digit from 0 to 9 consists to one model template separately. To legitimate the fact that our method does not need any scale normalisation either in space or in time, we create several samples of each model digits. Hence each sample has a different size in space and in time (according to the pronunciation speed of the subjects).

The probe sequences have a variable length, ranging from 4 to 10 digits in duration. The sorting of the digits can be either in an increasing order, in a decreasing order or at random. As for the model digit templates, each digit can have a different size in space and duration over time.

4.1 Comparative evaluation

We evaluate our approach by comparing it with two other different representations, audio and video only. Figure 2 shows the confusion matrices for the two different methods with different noise level. With video only, the model digit 8 is not correctly localised. This error is due to the movements of the lip for the digit 8 to be, in general, extremely limited. Consequently the matching is spread between every digits. With audio only, the accuracy of recognition rapidly decreases towards the point of complete failure as the quality of audio worsens (signal-to-noise ratio (SNR) decreases too). This observation motivates the integration between the audio and visual modalities to improve the recognition is any situation.

Indeed, Figure 3 shows that video alone can do a better job compared to poor audio alone. Moreover we observe that video plus noisy audio is better than poor audio alone. Finally, we can see that the combined model is significantly better than video alone in terms of both localisation accuracy and reduced ambiguity in selecting the right digit

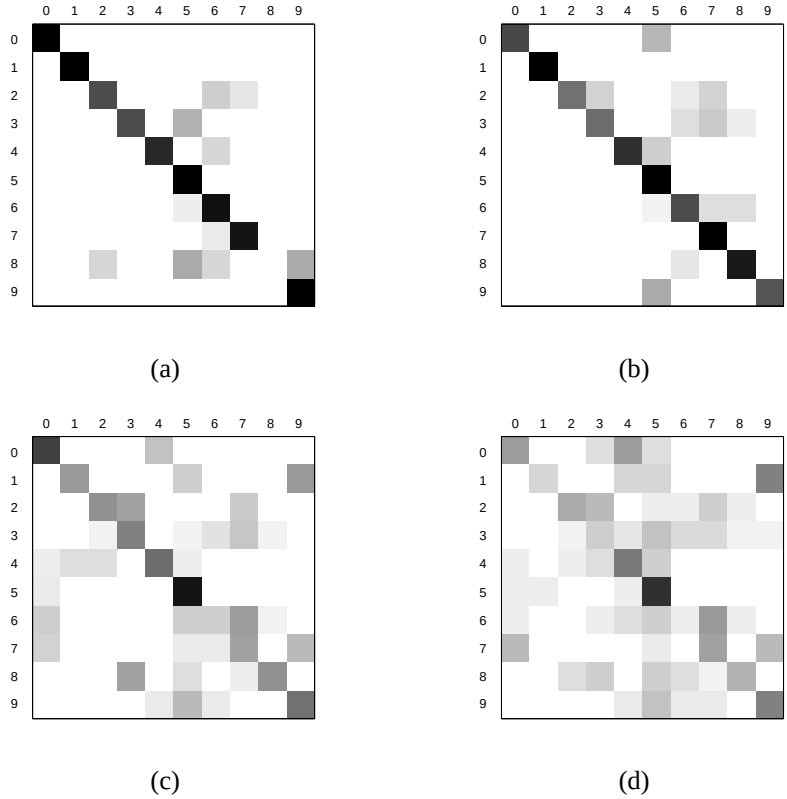


Figure 2: Confusion matrices from using (a) video only, (b) - (d) audio only with respectively a signal-to-noise ratio equal to 20db, 8db and 2db. The columns represent the model template indices whilst the rows account for correctness of digit recognition and localisation.

model template. Figure 4 shows several examples of experimental results with three different model templates on three different probe sequences. For those experiments, the speaker is male and the probe sequences contain 4 digits in duration. In each plot, we can see the probe sequence (shown by the biggest oblong), the macro-cuboïds with the highest probability to be in the probe sequence (the coloured oblongs) and the corresponding histogram. The biggest bin in each histogram indicates both the existence of the digit and its exact location in the probe sequence. In each probe sequence, the frame slices (4 slices per sequence) represent the first frame of a digit. Therefore the representation of the experiments is more apparent.

5 Conclusion

In this work we have shown the viability of a intermediate integration strategy. We address the problem of synchronisation between audio and visual signals. A linear interpolation

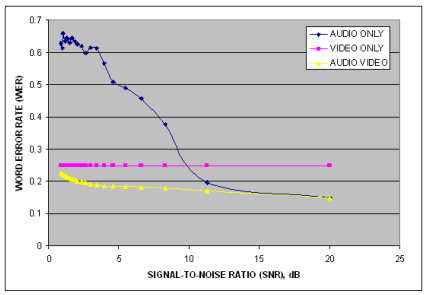
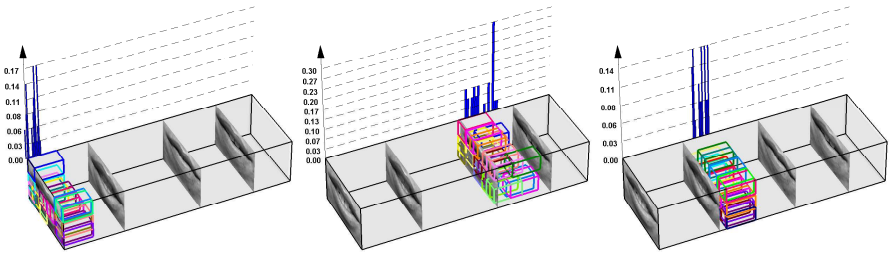


Figure 3: Comparison of audio-only, video-only and audio-visual ASR. A SNR of 20db means there is nearly no noise. 1db is a very noisy environment.



(a) (b) (c)

Figure 4: Experimental results using our approach. (a), (b) and (c) represent three different samples of the model digit 3 on the probe sequence with digits 3 , 1 , 5 and 9.

based approach adopted by existing techniques [9, 23] does not secure sufficient one-to-one correspondences and risks in inserting errors. Instead, we consider to construct a similar structure for both audio and visual feature selection and extraction.

Our systems consists of two major parts: (1) audio and visual feature selection and extraction, (2) intermediate fusion strategy by combining audio and video features in a probabilistic sequence matching function. Our first step defines and extracts automatically features extracted from 10 digit model templates and matches them into a probe video sequence. A model template for each digit is represented by a set of audio and visual features, using the same structure, which solves the synchronisation issue. Then a kernel-based maximum likelihood matching function is utilised to find the best match of all the audio-visual features candidates in a probe sequence for a model template.

The comparative evaluation between audio only, video only and audio-video ASR shows that our method outperform the other approaches. Experimental results demonstrate that the existence of a model digit and its exact location can be found in a probe sequence in every situation.

References

[1] P.S. Aleksic, J.J Williams, Zhilin Wu, and A.K. Katsaggelos. Audio-visual continuous speech recognition using MPEG-4 compliant visual features. In *ICIP*, 2002.

[2] O. Boiman and M. Irani. Similarity by composition. In *NIPS*, 2006.

[3] Yaron Caspi, Denis Simakov, and Michal Irani. Feature-based sequence-to-sequence matching. *IJCV*, 68(1):53–64, 2006.

[4] T. Chen. Audiovisual speech processing. *SPM*, 18(1):9–21, 2001.

[5] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. *PAMI*, 23(6):681–685, 2001.

[6] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VSPET*, 2005.

[7] Z. Fang, Z. Guoliang, and S. Zhanjiang. Comparison of different implementations of mfcc. *JCST*, 16(6):582–589, 2001.

[8] R. Göcke. Audio-video automatic speech recognition: an example of improved performance through multimodal sensor input. In *NICTA*, 2006.

[9] M.N. Kaynak, Qi Zhi, A.D. Cheok, K. Sengupta, and Ko Chi Chung. Audio-visual modeling for bimodal speech recognition. In *ICSMC*, 2001.

[10] M. Leszczynski and W. Skarbek. Viseme recognition - a comparative study. In *AVSS*, 2005.

[11] I. Matthews, T.F. Cootes, J.A. Bangham, S. Cox, and R. Harvey. Extraction of visual features for lipreading. *PAMI*, 24(2):198–213, 2002.

[12] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

[13] C.S. Myers and L.R. Rabiner. A comparative study of several dynamic time-warping algorithms for connected word recognition. *Bell System Tech. J.*, 60(7):1389–1408, 1991.

[14] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy. Dynamic bayesian networks for audio-visual speech recognition. *EURASIP*, 2002(11):1274–1288, 2002.

[15] J. C. Niebles, H. Wang, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In *BMVC*, 2006.

[16] S. Pachoud, S. Gong, and A. Cavallaro. Macro-cuboïd based probabilistic matching for lip-reading digits. In *CVPR*, 2008.

[17] E. Patterson, S. Gurbuz, Z. Tufekci, and J.N. Gowdy. Cuave: a new audio-visual database for multimodal human-computer interface research. In *ICASSP*, 2002.

[18] G. Potamianos, H.P. Graf, and E. Cosatto. An image transform approach for hmm based automatic lipreading. In *ICIP*, 1998.

[19] G. Potamianos, C. Neti, G. Iyengar, A. W. Senior, and A. Verma. A cascade visual front end for speaker independent automatic speechreading. *IJST*, 4:193–208, 2001.

[20] M.E. Sargin, Y. Yemez, E. Erzin, and A.M. Tekalp. Audiovisual synchronization and fusion using canonical correlation analysis. *IEEE Trans. Multimedia*, 9(7):1396–1403, 2007.

[21] W. H. Sumby and I. Pollack. Visual contribution to speech intelligibility in noise. *JASA*, 26(2):221–215, 1954.

[22] A. Valles, M. Gurban, and J. Thiran. Low-Dimensional Motion Features for Audio-Visual Speech Recognition. In *EUSIPCO*, 2007.

[23] W. Wang, D. Cosker, Y. Hicks, S. Saneit, and J. Chambers. Video assisted speech source separation. In *ICASSP*, 2005.